PHILOSOPHICAL TRANSACTIONS B

rstb.royalsocietypublishing.org

Opinion piece

Cite this article: Page RDM. 2016 DNA barcoding and taxonomy: dark taxa and dark texts. *Phil. Trans. R. Soc. B* **371**: 20150334. http://dx.doi.org/10.1098/rstb.2015.0334

Accepted: 10 February 2016

One contribution of 16 to a theme issue 'From DNA barcodes to biomes'.

Subject Areas:
taxonomy and systematics, bioinformatics

Keywords:
DNA barcoding, taxonomy, dark taxa, dark texts, digitization

Author for correspondence:
Roderic D. M. Page
e-mail: roderic.page@glasgow.ac.uk

DNA barcoding and taxonomy: dark taxa and dark texts

Roderic D. M. Page

Institute of Biodiversity, Animal Health and Comparative Medicine, College of Medical, Veterinary and Life Sciences, University of Glasgow, Glasgow G12 8QQ, UK

RDMP, 0000-0002-7101-9767

Both classical taxonomy and DNA barcoding are engaged in the task of digitizing the living world. Much of the taxonomic literature remains undigitized. The rise of open access publishing this century and the freeing of older literature from the shackles of copyright have greatly increased the online availability of taxonomic descriptions, but much of the literature of the mid- to late-twentieth century remains offline ('dark texts'). DNA barcoding is generating a wealth of computable data that in many ways are much easier to work with than classical taxonomic descriptions, but many of the sequences are not identified to species level. These 'dark taxa' hamper the classical method of integrating biodiversity data, using shared taxonomic names. Voucher specimens are a potential common currency of both the taxonomic literature and sequence databases, and could be used to help link names, literature and sequences. An obstacle to this approach is the lack of stable, resolvable specimen identifiers. The paper concludes with an appeal for a global 'digital dashboard' to assess the extent to which biodiversity data are available online.

This article is part of the themed issue 'From DNA barcodes to biomes'.

1. Introduction

As with many fields, digitization is having huge impact on the study of biodiversity. Museums and herbaria are engaged with turning physical, analogue specimens into digital objects, whether these are strings of As, Gs, Cs and Ts from DNA sequencing machines, or pixels obtained from a digital camera. Libraries and commercial publishers are converting physical books and articles into images, which are then converted into strings of letters using optical character recognition (OCR). Despite, sometimes, the acrimonious relationship between morphological and molecular taxonomy, there are striking parallels between the formation of DNA sequence databases in the twentieth century and the rise of natural history museums in the preceding centuries [1,2].

Viewed in this way, both classical taxonomy and genomics are in the business of digitizing life. Some of the challenges faced are similar, for example, algorithms developed for pairwise sequence alignment have applications in extracting articles from OCR text [3]. However, in other respects, the two fields are very different. Sequence data are approximately doubling every 18 months [4], whereas the number of new taxa described each year has remained essentially constant since the 1980s (see below). A challenge for sequence databases is how to handle exponential growth of data; for taxonomy, the challenge is often how to make a dent in the vast number of objects that do not have a digital representation [5]. This paper explores some of these issues, focusing on taxonomy and DNA barcoding.

2. Taxonomy

Among the many challenges faced by taxonomy is the difficulty of determining the size of the task it faces. Estimates of the number of species on Earth are uncertain and inconsistent, and show no signs of converging [6]. Some estimates,

THE ROYAL SOCIETY PUBLISHING

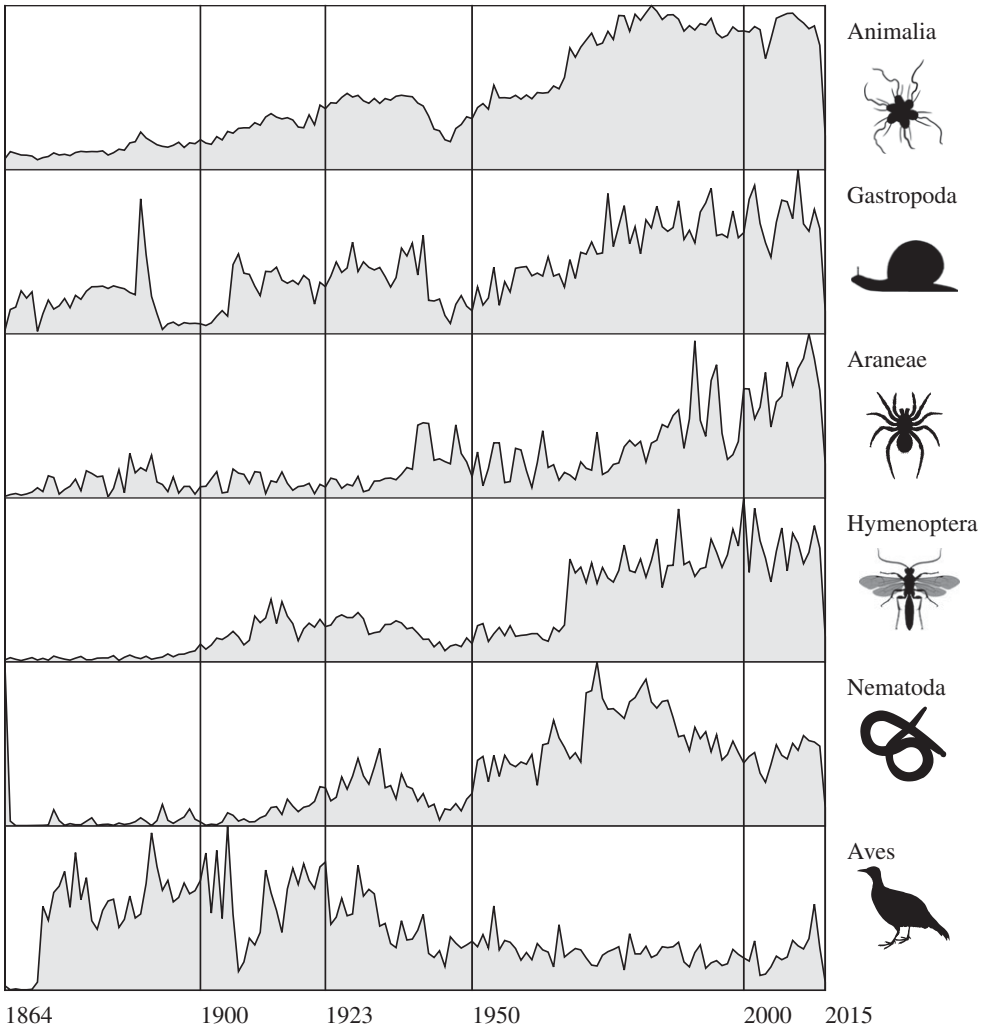

Figure 1. Trends of numbers of new names published each year for animals as a whole, and various taxonomic groups based on data in *Index to organism names* (ION) and BioNames (1923 is the year most published works became out of copyright in the USA). Animal pictures are from http://phylopic.org, and are either in the public domain or available under a Creative Commons CC-BY licence (Hymenoptera by Melissa Broussard; Nematoda by Michelle Site).

based on models of taxonomic effort, suggest that two-thirds of all species have already been described [7]. Analyses that use the number of authors per species description as a proxy for effort [8] ignore the global trend for an increasing number of authors per paper [9] and assume that the effort required per species description has remained constant over time. An alternative interpretation is that the quality of taxonomic description is increasing over time [10], reflecting both increased thoroughness and the availability of new technologies [11,12].

Rather than try and estimate an unknown (the number of species remaining to be described), here I focus on the current state of taxonomic knowledge. Given that we lack a comprehensive, global index of all species descriptions, discovering what we know about what we know is not entirely straightforward. For zoology, the nearest we have is the *Index to organism names* (ION, http://www.organismnames.com), which is based on *Zoological Record*. Figure 1 shows the numbers of new taxonomic names covered by the International Code on Zoological Nomenclature (animals plus some protozoan groups) that have been described each year based on data from ION, cleaned and augmented in BioNames (http://bionames.org) [13]. These data show an increase in overall numbers over time, with dips around the times of the two World Wars, followed by an essentially constant number each year since the mid-twentieth century. The pattern varies across taxa; some taxa show increasing numbers per year, but other taxonomic groups are essentially static or in decline, even in groups thought to be hyperdiverse such as nematodes [14].

(a) Digitizing the taxonomic literature

The rate of progress in biodiversity research is controlled by two factors, the speed with which we can discover and describe biodiversity, and the speed with which we can communicate that information [15]. Unlike most biological disciplines, the entire corpus of taxonomic literature since the mid-eighteenth century remains a vital resource for current-day research. In this way, taxonomy is similar to the digital humanities, which has not just 'big data' but 'long data' [16]. Not only is this because of the rules of nomenclature, which dictate (with some exceptions) that the name to use for a species is the oldest one published, it also reflects the uneven effort devoted to the study of different taxonomic groups [17]. For poorly known groups, the bulk of our knowledge of their biology may reside in the primary taxonomic literature.

Digitization is one step towards making taxonomic information available. Many commercial publishers have, on the face of it, done the taxonomic community a great service by digitizing whole back catalogues of relatively obscure journals. However, digitization is not the same as access, and many commercial publishers keep this scanned literature behind

paywalls. In some fields, legal issues around access have been side-stepped by constructing a 'shadow' dataset that summarizes key features of the data while still restricting access to the data itself. For example, by extracting phrases comprising a set of n words (n-grams) from Google Books, it is possible to create a dataset that contains valuable information without exposing the full text [18]. However, for taxonomic work, there does not seem to be an obvious way to extract a shadow. Agosti and co-workers [19,20] have explored ways to extract core facts from the literature and re-purpose these without violating copyright, though how many of their conclusions can be generalized across different national and international legal systems remains untested.

Apart from commercial digitization of the scientific literature, two other developments are accelerating access to taxonomic information. The first is the rise of open access publishing, notably journals such as *ZooKeys* that support sophisticated markup of the text [21]. This is increasing the number of recently described species that are published in a machine-readable form that can then be subject to further processing [22]. At the same time, the Biodiversity Heritage Library (BHL; http://biodiversitylibrary.org) [23] has embarked on large-scale digitization of legacy taxonomic literature. Although initially focusing on out of copyright literature (i.e. pre-1923 in the USA), BHL is increasingly getting permission from copyright holders to scan more recent literature as well. Coupled with tools such as BioStor (http://biostor.org) to locate and extract articles within the scanned volumes, BHL is fast becoming the largest available open access archive of biodiversity literature.

To quantify the extent to which the taxonomic literature has been digitized, for each decade, I counted the number of publications of new names in animals both with and without a digital identifier (such as a DOI, a PDF, a Handle or a URL to BioStor) in BioNames. The recent taxonomic literature is mostly digital: for the years 2010–2015, 60% of publications have a digital identifier, the bulk of these having a DOI. However, prior to the twenty-first century, more publications lack identifiers than have them, with the 1970s being the least digitized decade (figure 2).

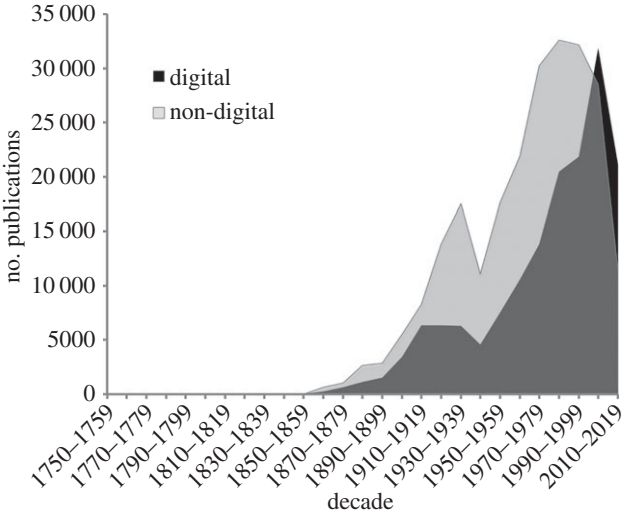

Figure 2. Number of taxonomic publications in BioNames for each decade, grouped by whether the publication has a digital identifier (e.g. a DOI, a link to JSTOR, BHL, BioStor, etc.). Publications containing new taxonomic names but lacking a digital identifier outnumber those that do have an identifier until 2000, represented here by the non-digital publication distribution (light grey) obscuring the digital distribution (black) until that date. The decline in both categories at the right of the chart reflects incomplete data for the current decade (2010–2020).

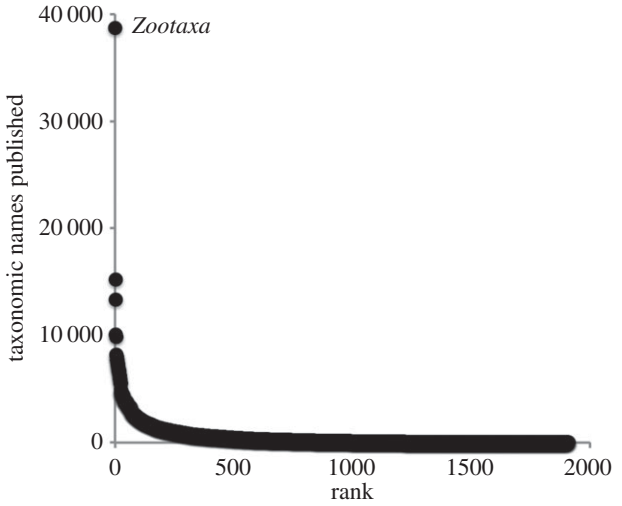

Figure 3. Number of taxonomic names published in each journal plotted against rank order for that journal. Note the distinctiveness of the first-ranked journal (*Zootaxa*). Data from BioNames.

(b) The long tail of taxonomic literature

Another challenge presented by the taxonomic literature is that it is highly decentralized, being spread across numerous journals (figure 3). What is striking is the dominance of animal taxonomy by the 'megajournal' *Zootaxa*, and yet this journal has published only 15% of the new names that have been minted since 2000. The taxonomic literature has a very 'long tail' of small, often obscure journals that contain a few taxonomic publications. Long tails require significant effort to index [24] although the *Zoological Record* claims 90% coverage of the taxonomic literature [25], in some taxa, there may be significantly greater gaps [26]. Conversely, if we set our sights lower, then long tail distributions mean that we can get a substantial fraction of the names from a small number of journals (the 'low hanging fruit'). Indeed, the first 20% of the journals in figure 3 contain 80% of the names in BioNames that are linked to a publication. Unfortunately, many of these journals are not currently available digitally.

The picture that emerges from our knowledge of the taxonomic literature is the recent literature is mostly digital, identified with DOIs, and some of it is open access. However, much of our fundamental knowledge of the world's biodiversity, particularly that published in the mid-to-late-twentieth century remains digitally inaccessible (figure 2). Between the twenty-first century trend towards digitization and open access and the removal of restrictions pre-1923 as copyright expires lies a great body of twentieth century work that will require considerable effort to make available.

(c) Genomics

In contrast with taxonomic knowledge, which is widely scattered, most genomic information is highly centralized, being stored in the three components of the International Nucleotide Sequence Database Collaboration (INSDC), namely GenBank, EMBL and the DDBJ [27]. Taxonomic name 'databases' more closely resemble digitized library catalogues, whereas sequence databases contain the actual sequences, which

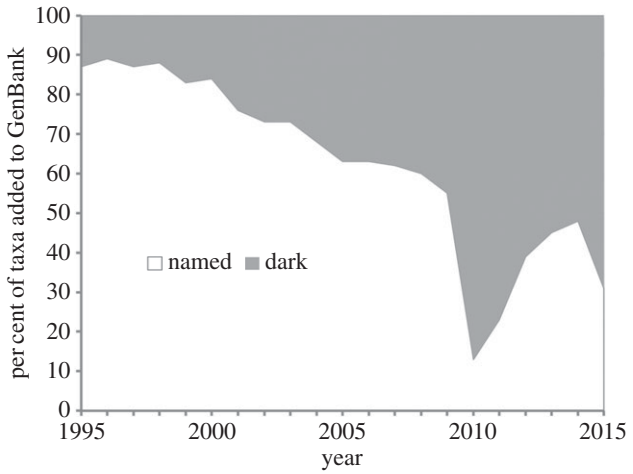

Figure 4. Growth of dark taxa in GenBank for invertebrate sequences. For each year, the graph shows the percentage of species-level 'invertebrate' taxa added during that year that do not have formal scientific names. The prominent drop in relative proportion of named taxa around 2010 is due to the addition of DNA barcodes from BOLD that lacked formal scientific names.

means we can compute over them. For example, a researcher with a new sequence can discover a lot about that sequence by a simple BLAST search [28], whereas a taxonomist armed only with a name will struggle to get computable data from the name alone.

Although the bulk of the world's sequence data are available in the INSDC, this is not the case for DNA barcodes, most of which reside in the Barcode of Life Data system (BOLD) [29]. Since 2009, BOLD has released some 2.5 million DNA barcodes, with updates every few months. Discovering how many of these barcodes are in GenBank is not entirely straightforward. Barcodes in GenBank may be flagged with the 'BARCODE' keyword (531 469 sequences at the time of writing), have a 'LinkOut' pointing to the BOLD database (60 684 sequences), or be listed the BioProject database [30] under accession PRJNA37833 (194 727 sequences). Because an individual sequence may meet one or more of these criteria, the sum total of sequences found by these searches (786 880) overestimates the total number of barcodes found by these methods. However, there are many barcode sequences that do not match any of these criteria. A dataset supplied by Sujeevan Ratnasingham lists 2 645 177 publicly available DNA barcodes in BOLD of which only half (1 317 132) have been shared with GenBank. The other half remain 'siloed' in BOLD.

(d) Dark taxa

As desirable as data sharing is, it is not without complications. In 2011, I coined the phrase 'dark taxa' (http://iphylo.blogspot.co.uk/2011/04/dark-taxa-genbank-in-post-taxonomic.html; see also [31]) to refer to species in GenBank that lacked formal scientific names. Typically, they will have a name that comprises a genus name and some combination of letters and numbers to make the name unique within GenBank (e.g. a specimen code or the first letter of the last names of the researchers that deposited the sequence). For this paper, I have updated the analysis to include sequences published up to the time of writing (figure 4).

The pattern shown in figure 4 likely reflects a combination of processes. If most of the taxa being added to GenBank represent species that have already been described, then the rate at which taxa can be identified (either by taxonomists or by researchers using their outputs, such as keys) is being outstripped by the pace of sequencing. Alternatively, dark taxa may represent unknown species, but we lack taxonomists capable of recognizing the taxa as new (and formally describing them). If taxonomic capacity is a limiting factor, then we would expect a gradual decline in percentage of named taxa, which is the background pattern in figure 4. The growth of dark taxa might also reflect changing practices of molecular workers, for example in DNA barcoding where large numbers of specimens are sequenced and deposited into GenBank labelled with specimen codes rather than taxonomic names. Indeed, the dramatic increase in the numbers of dark taxa in 2010 is mostly due to sequences from the BOLD project (recognized by taxa with the prefix 'BOLD') being added. Even if we allow for the import of unidentified BOLD sequences as a one-off event, at present less than half the newly sequenced invertebrate taxa being added to GenBank have been identified to species level. We have little idea whether these dark taxa represent newly discovered biodiversity, or are taxa that we already know about but have simply failed to link to already described species.

(e) Integrating biodiversity information

Typically, integration across biodiversity databases is achieved using taxonomic names [32], but the rise of dark taxa makes this problematic for an increasing fraction of sequence-based data. Even if we have names, these need not always mean the same thing [33]. As an example, figure 5*a* shows the distribution of the lizard *Morethia obscura* from the Global Biodiversity Information Facility (GBIF). For comparison, figure 5*b* shows a geophylogeny [34] for some DNA barcodes from BOLD for *Morethia obscura*, which reveals considerable phylogenetic structure within '*Morethia obscura*'. Specimens of this species are assigned several distinct Barcode Index Numbers (BINs) [35], implying that '*Morethia obscura*' comprises more than one species.

Although GBIF and BOLD present rather different views of the 'same' species, there is considerable overlap in the specimens used to construct figure 5*a*,*b*. For example, DNA barcode WAMMS012-10 was obtained from specimen WAMR127637, which also occurs in GBIF (as occurrence 691832269). Because the taxonomic concepts in GBIF and BOLD are explicitly defined with respect to sets of specimens, we can directly compare them, rather than rely on the possibly erroneous assumption that a given taxonomic name means the same thing in the two databases. Furthermore, as increasing numbers of type specimens are sequenced [36], we can more firmly associate names with sets of specimens, leading to a computable nomenclature where the name we assign to a set of specimens can be determined automatically [37]. Hence, our databases could be a lot more robust to the continual name changes that result from a nomenclatural system whereby taxonomic names are not 'opaque identifiers' but instead convey information about relationships (e.g. species sharing the same genus name are interpreted as being more closely related than those that do not).

Integrating databases using specimens is attractive, but not without its own set of problems. The biodiversity informatics community has yet to standardize identifiers for specimens, despite numerous efforts [38,39]; consequently, there may be little apparent overlap between specimen identifiers in

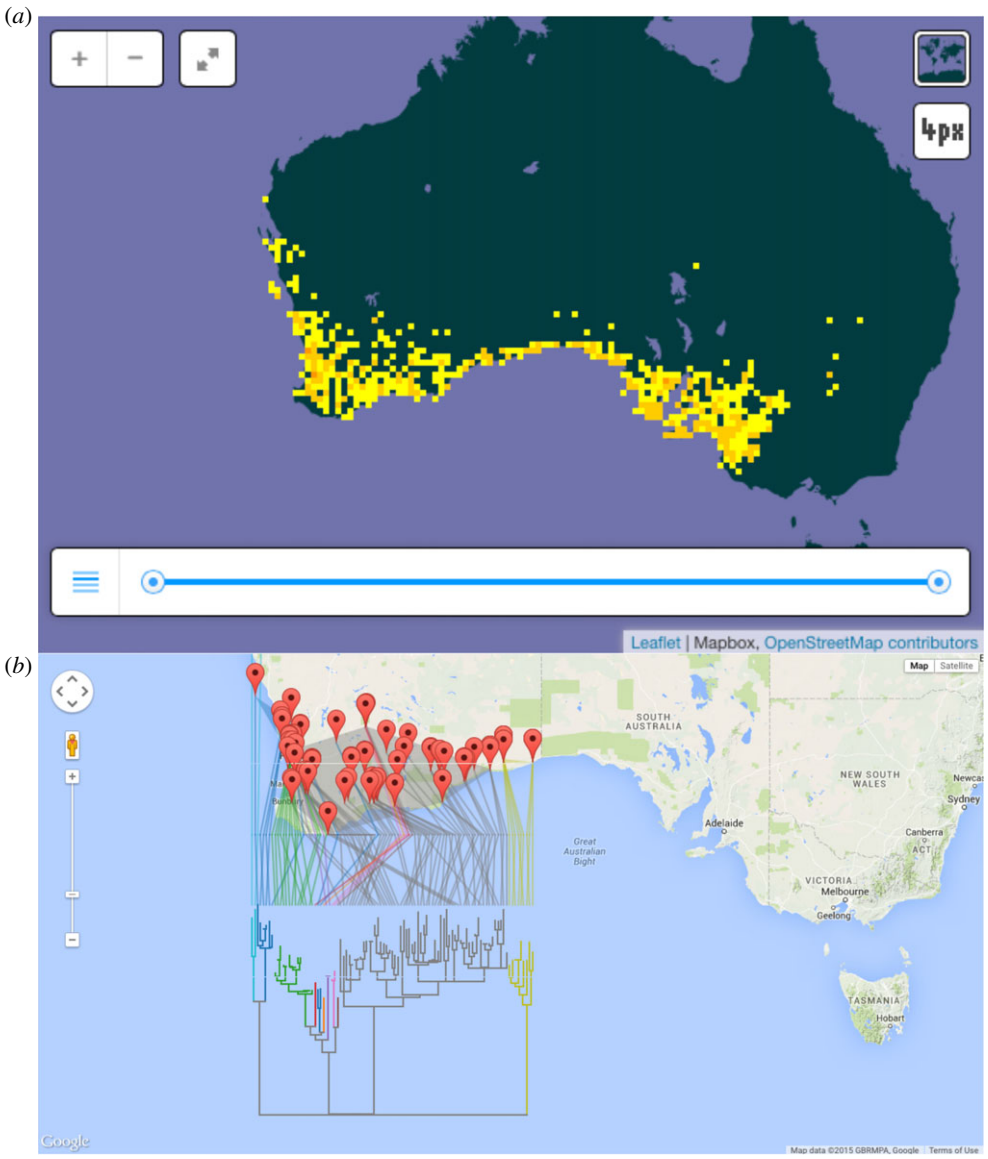

Figure 5. Comparison of *Morethia obscura* in GBIF (*a*) with DNA barcodes from the same taxon in BOLD (*b*). Note that the phylogeographic structure in the sequence data (which are assigned to several different BINs) implies the existence of multiple species within *Morethia obscura*.

different databases [40]. As an example, despite the limited sharing of data between BOLD and GBIF, there are already barcoded specimens in GBIF. To illustrate, consider the DNA barcode GWORH520-09 from sample 'BC ZSM Lep 10234'. GBIF does not have this record from BOLD, but it does have the specimen BC ZSM Lep 10234 provided by the host institution [41]. The DNA barcode from this specimen is also in GenBank, and because that record is georeferenced, it has been ingested by GBIF as part of the geographically tagged INSDC sequences dataset [42]. Hence, GBIF has duplicate records for this barcoded moth, neither provided directly by BOLD (figure 6). Merging and de-duplicating specimen-based records is going to be a significant challenge for global aggregators such as GBIF.

3. Summary

Both taxonomy and barcoding are actively digitizing the living world. The description of new animal taxa is essentially proceeding at a constant rate, generating a steadily growing legacy of taxonomic literature into which digitization has made modest inroads. In contrast, nucleotide sequence databases are growing exponentially. Nucleotide sequences are 'born digital' and readily computable; for example they can be clustered into BINs of similar sequences, or phylogenies of the type shown in figure 5. Given the obvious overlap between the goals of classical taxonomy and barcodes, the lack of digital overlap between these two endeavours is disconcerting. Many barcodes lack taxonomic names ('dark taxa'), and much of the primary taxonomic literature has not been digitized ('dark texts'). Integrating barcodes and taxonomy at scale is going to be significant challenge, as indeed will be integrating barcodes into mainstream sequence databases. Mapping between databases using taxonomic names seems the obvious approach, but the abundance of dark taxa shows this has not been entirely successful. Alternatives such as integration via specimens show promise, but are hampered by the lack of stable specimen identifiers. If we are to make progress the stubborn problem of the lack of unique, persistent identifiers, and crosslinks between those identifiers needs to be tackled in earnest [43,44].

As a postscript, in writing this opinion piece, I have had to write custom scripts to query various databases in an

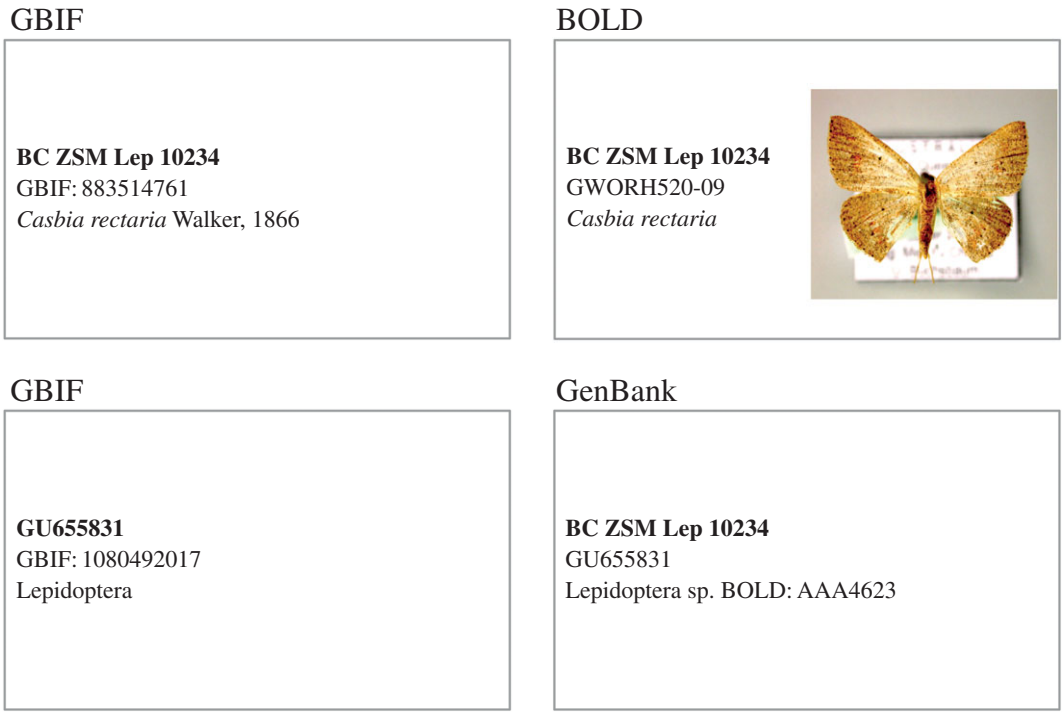

Figure 6. Illustration of multiple records for the same specimen of *Casbia rectaria* in GBIF. The voucher specimen for the DNA barcode is in GBIF (provided by the institution housing the specimen). The COI barcode sequence from this specimen is in both BOLD and GenBank, and because the sequence is georeferenced it is in GBIF as part of a dataset of georeferenced DNA sequences.

ad hoc manner (see http://github.com/rdmpage/dna-barcode-paper), trying to extract and assemble information that gives insight into the current state of biodiversity digitization. For these analyses and visualizations to have broader utility, it would be desirable to have some way of consistently and automatically doing these analyses, in effect creating a 'dashboard' of digitization that would enable us to not only see where we are as a field, but also suggest directions in which we could be heading. Many of the projects discussed in this article (mine included) use tools such as Google Analytics to provide detailed data on how users interact with their web sites [45]; it would be desirable to have similarly sophisticated tools to explore the actual data those sites are providing.

Data accessibility. Data and scripts used to create the figures are available from GitHub repository for this article at https://github.com/rdmpage/dna-barcode-paper.

Competing interests. I have no competing interests.

Funding. I received no funding for this study.

Acknowledgements. I thank Paul Hebert for the invitation to speak at the Sixth International Barcode of Life Conference. Some of the ideas discussed here were first developed on posts on my blog iPhylo (http://iphylo.blogspot.com) and benefited from feedback from people who left comments on those posts. Pete Hollingsworth and three anonymous reviewers provided very helpful critiques of the manuscript. Sujeevan Ratnasingham kindly provided a listing of DNA barcodes and their corresponding GenBank accession numbers.

References

1. Strasser BJ. 2008 GenBank–natural history in the 21st century? *Science* **322**, 537–538. (doi:10.1126/science.1163399)
2. Strasser BJ. 2011 The experimenter's museum: GenBank, natural history, and the moral economies of biomedicine. *Isis* **2011**, 60–96. (doi:10.1086/658657)
3. Page RDM. 2011 Extracting scientific articles from a large digital archive: BioStor and the Biodiversity Heritage Library. *BMC Bioinformatics* **12**, 187. (doi:10.1186/1471-2105-12-187)
4. Lathe W, Williams J, Mangan M, Karolchik D. 2008 Genomic data resources: challenges and promises. *Nat. Educ.* **1**, 2.
5. Ariño AH. 2010 Approaches to estimating the universe of natural history collections data. *Biodivers. Inf.* **7**. (doi:10.17161/bi.v7i2.3991)
6. Caley MJ, Fisher R, Mengersen K. 2014 Global species richness estimates have not converged. *Trends Ecol. Evol.* **29**, 187–188. (doi:10.1016/j.tree.2014.02.002)
7. Costello MJ, Wilson S, Houlding B. 2011 Predicting total global species richness using rates of species description and estimates of taxonomic effort. *Syst. Biol.* **61**, 871–883. (doi:10.1093/sysbio/syr080)
8. Joppa LN, Roberts DL, Pimm SL. 2011 The population ecology and social behaviour of taxonomists. *Trends Ecol. Evol.* **26**, 551–553. (doi:10.1016/j.tree.2011.07.010)
9. Aboukhalil R. 2014 The rising trend in authorship. *The winnower*. (doi:10.15200/winn.141832.26907)
10. Sangster G, Luksenburg JA. 2014 Declining rates of species described per taxonomist: slowdown of progress or a side-effect of improved quality in taxonomy? *Syst. Biol.* **64**, 144–151. (doi:10.1093/sysbio/syu069)
11. Stoev P *et al.* 2013 *Eupolybothrus cavernicolus* Komerički & Stoev sp. n. (Chilopoda: Lithobiomorpha: Lithobiidae): the first eukaryotic species description combining transcriptomic, DNA barcoding and micro-CT imaging data. *BDJ* **1**, e1013. (doi:10.3897/bdj.1.e1013)
12. Akkari N, Enghoff H, Metscher BD. 2015 A new dimension in documenting new species: high-detail imaging for myriapod taxonomy and first 3D cybertype of a new millipede species (Diplopoda, Julida, Julidae). *PLoS ONE* **10**, e0135243. (doi:10.1371/journal.pone.0135243)
13. Page RDM. 2013 BioNames: linking taxonomy, texts, and trees. *PeerJ* **1**, e190. (doi:10.7717/peerj.190)

14. Blaxter M. 2003 Molecular systematics: counting angels with DNA. *Nature* **421**, 122–124. (doi:10.1038/421122a)
15. Pentcheff ND. 2010 Copyrights and digitizing the systematic literature: the horror . . . the horror. . .. *Nat. Preced.* (doi:10.1038/npre.2010.4644)
16. Aiden E, Michel J-B. 2013 *Uncharted: big data as a lens on human culture*. New York, NY: Riverhead Books.
17. May RM. 1988 How many species are there on Earth? *Science* **241**, 1441–1449. (doi:10.1126/science.241.4872.1441)
18. Michel J-B *et al.* 2010 Quantitative analysis of culture using millions of digitized books. *Science* **331**, 176–182. (doi:10.1126/science.1199644)
19. Agosti D, Egloff W. 2009 Taxonomic information exchange and copyright: the Plazi approach. *BMC Res. Notes* **2**, 53. (doi:10.1186/1756-0500-2-53)
20. Patterson DJ, Egloff W, Agosti D, Eades D, Franz N, Hagedorn G, Rees JA, Remsen DP. 2014 Scientific names of organisms: attribution, rights, and licensing. *BMC Res. Notes* **7**, 79. (doi:10.1186/1756-0500-7-79)
21. Penev L *et al.* 2010 Semantic tagging of and semantic enhancements to systematics papers: ZooKeys working examples. *ZooKeys* **50**, 1–16. (doi:10.3897/zookeys.50.538)
22. Miller J *et al.* 2015 Integrating and visualizing primary data from prospective and legacy taxonomic literature. *Biodiversity Data J.* **3**, e5063. (doi:10.3897/bdj.3.e5063)
23. Gwinn NE, Rinaldo C. 2009 The Biodiversity Heritage Library: sharing biodiversity literature with the world. *IFLA J.* **35**, 25–34. (doi:10.1177/0340035208102032)
24. Edwards MA, Thorne MJ. 1993 Reply to 'Supraspecific names of molluscs: a quantitative review'. *Malacologia* **35**, 153–154.
25. Thorne J. 2003 *Zoological Record* and registration of new names in zoology. *Bull. Zool. Nomenclature* **60**, 7–11.
26. Bouchet P, Rocroi J-P. 1992 Supraspecific names of molluscs: a quantitative review. *Malacologia* **34**, 75–86.
27. Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. 2012 GenBank. *Nucleic Acids Res.* **41**, D36–D42. (doi:10.1093/nar/gks1195)
28. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990 Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410. (doi:10.1016/s0022-2836(05)80360-2)
29. Ratnasingham S, Hebert PDN. 2007 BOLD: the Barcode of Life Data System (http://www.barcodinglife.org). *Mol. Ecol. Notes* **7**, 355–364. (doi:10.1111/j.1471-8286.2007.01678.x)
30. Barrett T *et al.* 2011 BioProject and BioSample databases at NCBI: facilitating capture and organization of metadata. *Nucleic Acids Res.* **40**, D57–D63. (doi:10.1093/nar/gkr1163)
31. Parr CS, Guralnick R, Cellinese N, Page RDM. 2012 Evolutionary informatics: unifying knowledge about the diversity of life. *Trends Ecol. Evol.* **27**, 94–103. (doi:10.1016/j.tree.2011.11.001)
32. Patterson DJ, Cooper J, Kirk PM, Pyle RL, Remsen DP. 2010 Names are key to the big new biology. *Trends Ecol. Evol.* **25**, 686–691. (doi:10.1016/j.tree.2010.09.004)
33. Kennedy J. 2003 Supporting taxonomic names in cell and molecular biology databases. *OMICS J. Integr. Biol.* **7**, 13–16. (doi:10.1089/153623103322006508)
34. Page RDM. 2015 Visualising geophylogenies in web maps using GeoJSON. *PLoS Curr.* (doi:10.1371/currents.tol.8f3c6526c49b136b98ec28e00b570a1e)
35. Ratnasingham S, Hebert PDN. 2013 A DNA-based registry for all animal species: the Barcode Index Number (BIN) system. *PLoS ONE* **8**, e66213. (doi:10.1371/journal.pone.0066213)
36. Federhen S. 2014 Type material in the NCBI taxonomy database. *Nucleic Acids Res.* **43**, D1086–D1098. (doi:10.1093/nar/gku1127)
37. Pullan MR, Watson MF, Kennedy JB, Raguenaud C, Hyam R, Raguenaud C. 2000 The Prometheus Taxonomic Model: a practical approach to representing multiple classifications. *Taxon* **49**, 55. (doi:10.2307/1223932)
38. Guralnick RP *et al.* 2015 Community next steps for making globally unique identifiers work for biocollections data. *ZooKeys* **494**, 133–154. (doi:10.3897/zookeys.494.9352)
39. Hyam R, Drinkwater RE, Harris DJ. 2012 Stable citations for herbarium specimens on the internet: an illustration from a taxonomic revision of *Duboscia* (Malvaceae). *Phytotaxa* **73**, 17–30. (doi:10.11646/phytotaxa.73.1.4)
40. Guralnick R, Conlin T, Deck J, Stucky BJ, Cellinese N. 2014 The trouble with triplets in biodiversity informatics: a data-driven case against current identifier practices. *PLoS ONE* **9**, e114069. (doi:10.1371/journal.pone.0114069)
41. GBIF. 2015 Zoologische Staatssammlung Muenchen - International Barcode of Life (iBOL) - Barcode of Life Project specimen data. (doi:10.15468/tfpnkp)
42. GBIF. 2014 Geographically tagged INSDC sequences. (doi:10.15468/cndomv)
43. Page RDM. 2008 Biodiversity informatics: the challenge of linking data and the role of shared identifiers. *Brief. Bioinformatics* **9**, 345–354. (doi:10.1093/bib/bbn022)
44. Page RDM. 2016 Surfacing the deep data of taxonomy. *ZooKeys* **550**, 247–260. (doi:10.3897/zookeys.550.9293)
45. Jones T, Baxter D, Hagedorn G, Legler B, Gilbert E, Thiele K, Vargas-Rodriguez Y, Urbatsch L. 2014 Trends in access of plant biodiversity data revealed by Google Analytics. *Biodiversity Data J.* **2**, e1558. (doi:10.3897/bdj.2.e1558)